自然言語処理 95－7
(1993. 5. 20)

単語の結束性にもとづいてテキストを場面に分割する試み

小嶋 秀樹
電気通信大学 大学院
情報工学専攻
〒182 東京都 調布市 調布ヶ丘 1-5-1
E-mail. xkozima@phaeton.cs.uec.ac.jp

古郡 廷治
電気通信大学
情報工学科
〒182 東京都 調布市 調布ヶ丘 1-5-1
Tel. 0424-83-2161 (ex.4461)

あらまし 本論文では, テキスト区画 (とくに英語の物語における場面) の境界を推定するための統計的な指標として, LCP (lexical cohesion profile) を提案する. テキスト区画は, 意味的に一貫した部分テキストであり, そこに現われる単語がたがいに結束性 (lexical cohesion) によって結ばれる傾向をもつ. LCP は, テキスト上を移動する一定幅の窓から見える単語列の結束度を記録したものである. 単語列の結束度は, 英語辞書から規則的に構成された意味ネットワーク上の活性伝播によって計算される. 人間の直感による場面分割と比較した結果, LCP の変化が場面境界とつよい相関をもつことが確かめられた. LCP によって推定される場面境界は, 照応や省略を解決するための手がかりとして利用できる.

和文キーワード テキスト構造, テキスト分割, 結束性, 一貫性, 意味ネットワーク

Text Segmentation Based on Lexical Cohesion

Hideki Kozima
Course in Computer Science
and Information Mathematics,
Graduate School,
University of Electro-Communications
1-5-1, Chofugaoka, Chofu,
Tokyo 182, Japan
(xkozima@phaeton.cs.uec.ac.jp)

Teiji Furugori
Department of Computer Science
and Information Mathematics,
University of Electro-Communications
1-5-1, Chofugaoka, Chofu,
Tokyo 182, Japan
Tel. +81-424-83-2161 (ex.4461)

Abstract This paper proposes a new indicator of text segment, called LCP (lexical cohesion profile), based on lexical cohesion between words. A text segment is a coherent scene in which the words tend to have lexical cohesion with each other. LCP is a record of cohesiveness of words in an interval moving on the text. The cohesiveness is computed by spreading activation on a semantic network constructed systematically from an English dictionary. Comparison with the segment boundaries marked by a number of subjects shows that LCP closely correlates with the human intuition. LCP provides valuable information for resolving anaphora and ellipsis.

英文 key words Text Structure, Segmentation, Lexical Cohesion, Coherence, Semantic Network

1 はじめに

テキストは，単語や文の単なるあつまりではなく，何らかの目的[1]のために組織化されたテキスト構造をもっている．個々の単語や文は，それらをとりまく構造単位のなかで役割を与えられ，はじめて意味をもつ．テキスト構造は，個々の単語や文の意味を決定するための，また照応や省略の解決・曖昧性の解消などのための，文脈（制約・選好など）をもたらす．

テキスト区画（text segment）は，このような文脈に大きくかかわる構造単位であり，ある目的[2]をもって書かれた意味的なまとまりをもつ部分テキストである．[Grosz and Sidner, 1986] テキスト表層に明示された句読点や段落分けなどは，テキスト区画の境界を示唆するが，意味的にまとまった部分に分割するとはかぎらない．

たとえば物語におけるテキスト区画は，映画や劇における場面（scene）に相当し，そこでは同じ対象（人物・事物など）がほぼ同じ状況（場所・時刻・視点など）のもとで語られる．テキスト区画は，このような一貫性（coherence）をもち，そこに現われる単語がたがいに結束性（lexical cohesion）によって結ばれる傾向をもつ．[Morris and Hirst, 1991]

本論文では，テキスト区画（とくに英語の物語における場面）の境界を単語の結束性にもとづいてとらえるための統計的な指標，LCP（lexical cohesion profile）を提案する．LCP は，テキストの各位置について，その近傍（たとえば前後25語の範囲）にある単語列の結束度（cohesiveness）を記録したものである．単語列の結束度は，英語辞書から規則的に構成された意味ネットワーク上の活性伝播によって計算される．

このようにして得られた LCP は，意味的に異なる2つの場面の境界で極小値をとる．なぜなら，境界近傍の単語列は，意味的に異なる2つの部分列からなり，その全体としての結束度が低くなるからである．

以下，次節では，テキスト区画と結束性についての関連研究を概観し，本研究のとる方略を明らかにする．第3節では，意味ネットワークの構成方法と，活性伝播による結束度の計算方法を説明する．第4節では，LCP の計算方法と場面境界の推定方法を説明する．第5節では，LCP と人間の直感による場面分割とを比較し，LCP の妥当性を検証する．第6節では，LCP の可能性と問題点をまとめ，今後の研究を展望する．

2 テキスト区画と結束性

テキスト区画は，書き手の意図（つまりテキストの目的）によって意味的にまとめられた部分テキストである．読み手がテキストを理解するには，意味的な一貫性を手がかりに，このテキスト区画を認識しなければならない．したがって，どんな読み手もテキスト区画をほぼ同じように — つまり書き手によって意図されたように — とらえる傾向をもつ．

2.1 テキスト区画をとらえる研究

テキスト区画をとらえるいくつかの方法が，テキスト構造に関する研究のなかで提案されてきた．たとえば，手がかり語句（cue phrase）[Grosz and Sidner, 1986] は，テキスト区画の変化を示唆する．[3] しかしながら，手がかり語句を含めた従来の指標は，テキスト区画の意味的な一貫性を直接反映するものではなく，むしろ（おもに会話における）意味伝達を円滑にするための補助的な信号である．

Youmans [1991] は，テキスト区画の意味的な一貫性を直接とらえる指標として VMP（vocabulary-management profile）を提案した．VMP は，テキストの各位置について，その近傍の単語列に含まれる新規語彙数 — はじめてそのテキストに現われる単語の数 — の割合を記録したものである．そのグラフは情報の満ち引きを表わす周期的な山と谷をもち，その谷はテキスト区画の境界を示唆するという．

Youmans の VMP は書かれたテキストを分割するためのシンプルで客観的な指標であるが，我々が VMP を再現・検証した結果，つぎのような問題点が明らかになった:

---

[1] 読み手に効率よく知識を伝えること，読み手を楽しませること・納得または感動させること，書き手に対するよい評価を読み手にもたせること，など．[Grosz and Sidner, 1986; Schank, 1990]

[2] テキスト全体としての目的が異なれば — つまり，物語・論文といったジャンルごとに — テキスト区画がもちうる目的も異なる．

[3] たとえば，文頭の "By the way"，"One day" などは，新しいテキスト区画のはじまりを示唆する．

- 語彙密度 (総語彙数と総単語数の比) の高いテキストを正しく分割することができない.
- テキストの先頭部分 (約 250 語より前) や, テキストの後尾部分 (数千語より後) の VMP は, テキスト区画との相関がよわい.

これらの問題点は, VMP が単語の反復という現象だけにもとづいてテキスト区画を推定していることが原因である.

テキスト区画の各単語は, 表層的な反復としてだけでなく, 内容的な意味関係 (semantic relation) としても結束性をもっている. [Halliday and Hasan, 1976] この意味関係を含めた結束性をとらえなければ, テキスト区画の境界を正しく推定することはできない.

2.2 結束性をとらえる研究

単語間の結束性, とくに意味関係をとらえる方法として, Morris and Hirst [1991] はシソーラスの利用を提案している. この方法では, 2つの単語間に意味関係があることは, それらが Roget のシソーラス上で同じ (または関連した) クラスに含まれることと定義される.

この方法によって, 感情的・時系列的なものをのぞくほとんどの意味関係をとらえることができる:

- 同義語・上位語による同一参照 (coreference) (ex. `cat/pet`, `wine/alcohol`, etc.),
- 体系的 (systematic) な意味関係 (ex. `buy/sell`, `north/east`, etc.),
- 非体系的 (non-systematic) な意味関係 (ex. `waiter/restaurant`, `blood/red`, etc.).

しかしながら, この方法では, 単語間に結束性があるかないかの判別しかできず, 結束性のつよさを知ることができない.

テキスト区画に現われる個々の単語は, それぞれの役割をもって, それぞれの度合いで, その区画全体の一貫性に貢献している. テキスト区画の境界を正しく推定するには, 単語間の結束性のつよさ (つまり結束度) をとらえることが必要となる.

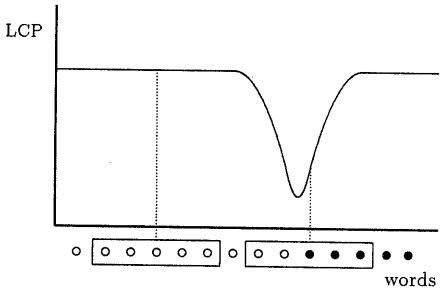


図 1. テキスト区画の境界と LCP の関係

2.3 テキスト区画の境界を推定する方略

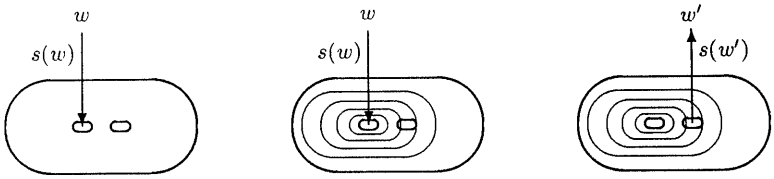
一貫性によって定義されるテキスト区画の境界をとらえるには, テキストの局所的な一貫性の度合いを推定する必要がある. テキストの一貫性を単語間の結束性によって推定できる [Morris and Hirst, 1991] のだから, 単語の反復にもとづく Youmans [1991] の統計的な方法に, 単語の結束度 (とくに意味関係のつよさ) を取り込めばよい.

ここで提案する LCP は, テキストの各位置について, その近傍の単語列の結束度 — 各単語相互の結束度 — を記録したものであり, テキストの各位置における一貫性の指標となる. 図 1 は, 横軸にテキスト上の位置, 縦軸に LCP をとって, LCP をグラフにしたものである. 意味的に異なる 2 つのテキスト区画 ("o o ··· o" と "● ● ··· ●") の境界では, その近傍の単語列が両方のテキスト区画を半分ずつ含むため, LCP は極小値をとることがわかる.

単語列の結束度は, 意味ネットワーク上の活性伝播 [Waltz and Pollack, 1985] によって計算される. 我々は, 英語辞書から規則的に意味ネットワークを構成し, 単語間の結束度および単語列の結束度を [0,1] の連続量として客観的に測定する方法を開発した. [小嶋・古郡, 1993] 次節では, 意味ネットワークの構成方法と結束度の測定方法について説明する.

3 結束度の計算

単語間の結束度は, 意味ネットワーク *Paradigme* における活性伝播によって, [0,1] の連続量として計算される. *Paradigme* は, 英語辞書 LDOCE (*Longman*



(1) 単語 w の活性化をはじめる. (2) 活性パターンが生成される. (3) 単語 w' の活性度を観測する.

図2. 単語間の類似度 $\sigma(w, w')$ の計算

Dictionary of Contemporary English) [1987] の特殊なサブセットから規則的に構成される.

Paradigme の構成方法と, 結束度の計算方法の詳細については, [小嶋・古郡, 1993] を参照されたい.[4]

3.1 意味ネットワーク Paradigme

ここで使用する辞書 LDOCE はつぎのような特徴をもつ英語辞書である — その約 56000 語の見出し語のすべてが, *Longman Defining Vocabulary* (以下 LDV) という 2851 語[5]の制限語彙 (とその派生語) だけを使って定義されている. LDV は単語の使用頻度調査 [West, 1953] などにもとづいて選ばれている.

意味ネットワーク *Paradigme* は, LDOCE から LDV を見出し語とするような各項目を節点に写像したものであり, 全部で 2851 の節点からなる. 各節点は, LDOCE の語義定義から規則的に生成されたリンク集合をもち, *Paradigme* 全体で 295914 のリンクをもっている. 各リンクの重みは, 節点の参照頻度などから規則的に計算される.

3.2 単語間の結束度の計算

Paradigme のある節点を一定時間にわたって活性化させると, その活性がリンクをとおして *Paradigme* 上に拡がり, 活性パターンが生成される. 単語 w, w' 間の結束度 $\sigma(w, w') \in [0, 1]$ は, この活性パターンを利用して, つぎのように計算される (図2参照):

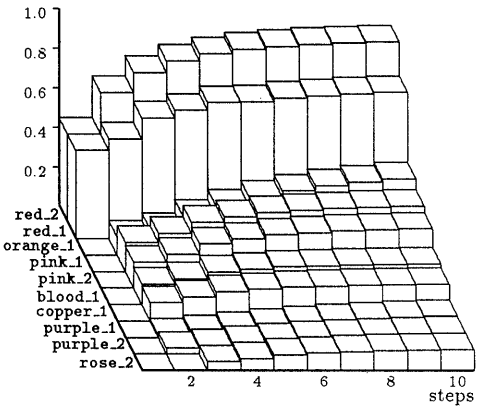


図3. 活性パターンの生成例

(単語 red を活性化させたとき, $t = 10$ で最も高い活性度をもつ 10 節点についての活性度の時間変化)

1. 単語 w をつよさ $s(w)$ で 10 ステップのあいだ活性化させる.
2. この結果, 図3のような活性パターン $P(w)$ が生成される.
3. $P(w)$ における w' の活性度 $a(P(w), w')$ を観測する. $\sigma(w, w')$ は $s(w') \cdot a(P(w), w')$ となる.

単語 w の重要度 $s(w)$ は, West [1953] のコーパスにおける w の情報量を [0,1] に正規化したものである.

この方法によって, LDV (およびその派生語) に含まれる任意の単語間の結束度を計算することができる. 単語 w, w' 間の結束度 $\sigma(w, w')$ は, 単語間の意味的なつながりのつよさに応じて高くなる:

$$\begin{aligned}
\sigma(\texttt{cat}, \texttt{pet}) &= 0.133722, \\
\sigma(\texttt{cat}, \texttt{hat}) &= 0.001784, \\
\sigma(\texttt{buy}, \texttt{sell}) &= 0.135686, \\
\sigma(\texttt{buy}, \texttt{walk}) &= 0.007993,
\end{aligned}$$

[4] [小嶋・古郡, 1993] では, ここでいう結束性を「意味的な類似度」とよんでいる.

[5] LDV は約 2200 語彙であるが, 小項目辞書である LDOCE の見出し語数に換算すると 2851 語となる.

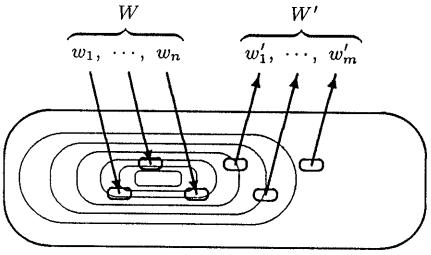


図4．単語列間の結束度の計算

σ (waiter, restaurant) = 0.175699 ,
σ (painter, restaurant) = 0.006260 .

結束度 σ には方向性があり，一般に $\sigma(w, w') \neq \sigma(w', w)$ となることに注意されたい．

意味の豊かな（重要度の高い）単語は高い結束度をとり，機能語などの意味の薄い（重要度の低い）単語は低い結束度をとる：

σ (red, blood) = 0.111443 ,
σ (of, blood) = 0.001041 .

また，自己結束度 $\sigma(w, w)$ が 1 にならず，単語の重要度 $s(w)$ に依存することに注意されたい：

σ (waiter, waiter) = 0.596803 ,
σ (of, of) = 0.045256 .

3.3 LDV に含まれない単語の結束度の計算

LDV に含まれない単語の結束度は，その単語を LDOCE によって語義 — これは LDV だけからなる単語列である — に言い換えることによって，単語列間の結束度として計算することができる．(図4参照)

単語列 $W = \{w_1, \cdots, w_n\}$ と $W' = \{w'_1, \cdots, w'_n\}$ の間の結束度 $\sigma(W, W')$ は，つぎのように計算される：

$$\sigma(W, W') = \psi\left(\sum_{w' \in W'} s(w') \cdot a(P(W), w')\right) .$$

$P(W)$ は，各単語 $w_i \in W$ をつよさ $s(w_i)^2 / \sum s(w_k)$ で 10 ステップのあいだ活性化させることによって生成される活性パターンであり，$a(P(W), w')$ は，この活性パターンにおける単語 w' の活性度である．また，ψ は σ の値を [0,1] に制限する関数である．

たとえば，LDV に含まれない単語 linguistics と stylistics の結束度は，つぎのように計算される：

σ(linguistics, stylistics)
= σ({the, study, of, language, in, general, and, of, particular, languages, and, their, structure, and, grammar, and, history },
{the, study, of, style, in, written, or, spoken, language })
= 0.140089 .

この例はどちらも LDV に含まれない場合であるが，LDV に含まれる単語をただ1つの単語からなる単語列として扱えば，このほかの場合もおなじように計算できる．

この方法によって，直接・間接のちがいはあるが，LDOCE の見出し語（およびその派生語）に含まれる任意の単語間の結束度を計算することができる．

4 LCP (Lexical Cohesion Profile)

LCP は，テキストの各位置（第 i 語）における近傍の単語列 S_i の結束度 $c(S_i)$ を記録したものである．これはちょうど，テキスト上を移動する一定幅（たとえば 51 語）の窓から見える単語列の結束度を計算し，それを窓の中心位置の LCP の値としたものである．

4.1 単語列の結束度の計算

単語列 $S = \{w_1, \cdots, w_n\}$ の結束度 $c(S)$ は，**3.3** で説明した単語列間の結束度 σ を利用して，つぎのように定義される：

$$c(S) = \sigma(S, S) = \psi\left(\sum_{w \in S} s(w) \cdot a(P(S), w)\right) .$$

単語列 S から生成した活性パターン $P(S)$ は，各単語 $w \in S$ の意味的な重ねあわせであり，S の全体的・平均的な意味を表わしている．たとえば，単語列 {red, alcoholic, drink} から生成される活性パターンは，図5のようになる．直接に活性化される単語（上位5つの節点）につづいて，bottle や wine といった単語もつよく活性化していることに注目されたい．

単語列 S の結束度 $c(S)$ は，このような活性パターン $P(S)$ における S の活性度であり，S の意味的な一様度（または，ひずみ[6]の少なさ）を表わしている．なぜなら，$c(S)$ の定義は，各単語 $w \in S$ が，S の全体的・

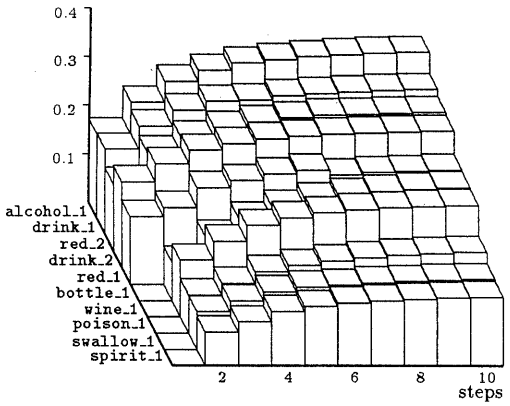


図5． 単語列から活性パターンを生成した例
(単語列 {red, alcoholic, drink} から生成した場合)

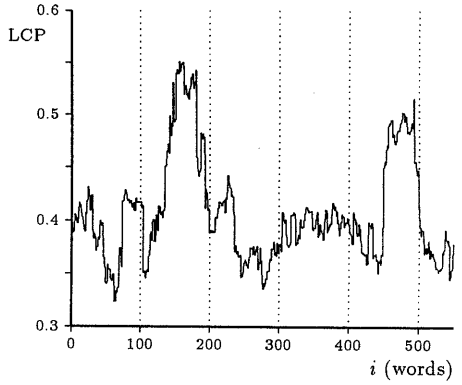


図6． LCP の計算例 (窓関数なし)

平均的な意味 $P(S)$ にどれくらい結束しているかを表わしているからである.

単語列 S の結束度 $c(S)$ は, S の意味的な一貫性の度合いを示唆する. たとえば, 3つの文からなる意味的に一貫したテキストと, 辞書 LDOCE から例文を3つランダムに取り出したテキストは, それぞれつぎのような結束度をもつ:

c("Molly saw a cat. It was her family pet. She wished to keep a lion."
= 0.403239 (一貫性あり),
c("Put on your clothes at once. I can not walk ten miles. There is no one here but me.")
= 0.250840 (一貫性なし).

4.2 LCP の計算

単語列とみなしたテキスト $T=\{w_1,\cdots,w_N\}$ (句読点などは無視) について, 位置 i の近傍の単語列 S_i をつぎのように定義する:

$$S_i=\{w_l,w_{l+1},\cdots,w_{i-1},w_i,w_{i+1},\cdots,w_{r-1},w_r\},$$
$$l = i-\Delta \quad (\text{if } i\leq\Delta, \text{ then } l=1),$$
$$r = i+\Delta \quad (\text{if } i>N-\Delta, \text{ then } r=N).$$

S_i は, 第 i 語が中心となるように置いた窓から見える単語列である. Δ は窓の幅をきめる定数であり, その長さはテキストの最初と最後の Δ 語をのぞいて $2\Delta+1$ となる.

LCP は, テキスト上の各位置 i について結束度 $c(S_i)$ を記録したものである. 図6は, ある短編小説について計算した LCP を, 横軸に i, 縦軸に $c(S_i)$ をとってグラフにしたものである. (ただし, 最初の 550 語のみ表示.) グラフの大きな谷 — たとえば $i=65$, 105, 200, 275, 440 など — は, このテキストの場面境界を示唆する. なぜなら, S_i が場面の内部にあるかぎりその結束度 $c(S_i)$ は高く維持されるが, S_i が2つの場面にまたがるとき $c(S_i)$ は低くなるためである. (図1参照)

しかしながら, 図6の LCP のグラフは細かく振動しているため, どの極小点を場面境界とみなせばよいか曖昧である. 場面境界を正しく推定するには, LCP の巨視的な変化をとらえることが必要である.

4.3 LCP の巨視的な特徴の明確化

LCP の巨視的な変化をより明らかにし, 場面境界をよりわかりやすくするために, 窓の幅と形を調節する.

窓の幅は, LCP の分解能に直接影響する. 窓が狭いほどより小さな場面の変化をとらえることができ, 窓が広いほど LCP の巨視的な特徴をより明らかにすることができる. 5 から 60 の間の 18 種類の Δ について比較した結果, 図6の LCP で使用した $\Delta=25$ の窓がもっともよい結果をもたらすことがわかった.

窓の形も, LCP の巨視的な特徴に影響する. そこで, **4.1** で説明した単語列 $S=\{w_1,\cdots,w_n\}$ から

[6] この「ひずみ」は, クラスタリングでいう distortion にあたる. $P(S)$ をセントロイド, σ の逆数を単語間の距離と考えればよい.

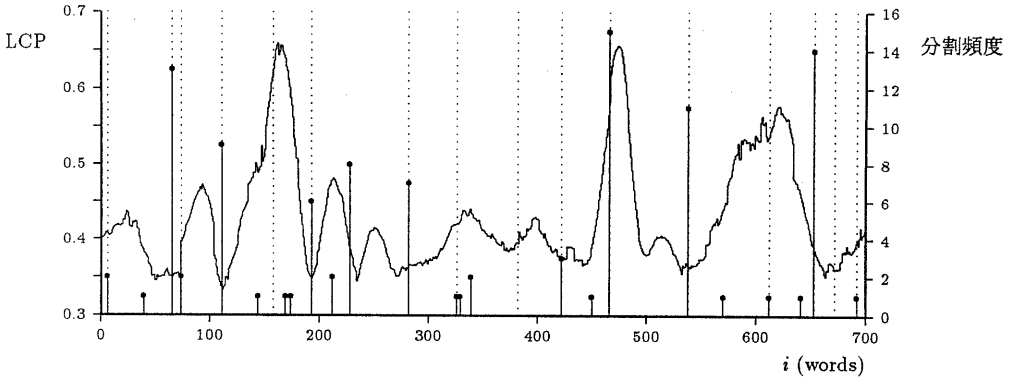


図7． LCP と人間の直観による場面分割との比較

(テキスト: *Springtime à la Carte* (最初の 700 語についてのみ表示), 窓関数: $\Delta = 25$ のハニング窓.)

活性パターン $P(S)$ を生成する手続きを，つぎのように拡張する — 各単語 $w_j \in S$ を，つよさ $M(j) \cdot s(w_j)^2 / \sum_j M(j) \cdot s(w_j)$ で 10 ステップのあいだ活性化させることによって $P(S)$ を生成する．

窓関数 $M(j)$ は，窓から見える単語列 S の先頭から j 番目の単語 w_j にかける重みを決定する．方形窓 (rectangular window, $M(j) = 1$)[7], 三角窓 (triangle window, $M(j) = 1 - |j - (\Delta+1)|/\Delta$) などを含む 6 種類の窓関数について比較した結果，ハニング窓:

$$M(j) = \frac{1}{2}\left(1 + cos\left(\frac{|j-(\Delta+1)|\pi}{\Delta}\right)\right),$$

がもっともよい結果をもたらすことがわかった．

次節で LCP と人間の直観による場面分割とを比較するが，そこでは $\Delta = 25$ (窓の幅は 51 語) のハニング窓を使って LCP を計算している．

5 LCP の検証

LCP の妥当性を検証するため，O.Henry の短編小説 *Springtime à la Carte* [Thornley, 1960] について計算した LCP と人間の直感によってとらえられた場面境界とを比較する．このテキストは，1620 語 (484 語彙)[8] であり，単語の使用頻度調査 [West, 1953] にもとづく 2000 語の制限語彙を使って書かれている．

まず，人間の直感による場面分割を観察する実験を行なった．16 人の被験者に，もとの段落構造の失われた 1 行 1 文のテキストを読ませ，文の切れ目 (110 箇所) のなかから場面境界と思われるものをいくつでも選択させた．のべ 214 箇所 (平均 13.38 箇所 / 人)，タイプ数 50 の場面境界が報告された．これをヒストグラムにした図 7 の棒グラフから，被験者がほぼ同じように場面境界をとらえ，その多くが点線で示された段落境界と一致することがわかる．

つぎに，同じテキストについて ($\Delta = 25$ のハニング窓による) LCP を計算し，図 7 のヒストグラムにに重ねた．1/3 以上の被験者に共通した場面境界 — これを共通場面境界とよぶことにする — の多くが LCP の巨視的な谷 (極小点) と一致しているのがわかる．たとえば $i = 192$ の極小点は，共通場面境界と (さらに段落境界とも) 一致している．この付近の部分テキスト ($i = 157 \sim 227$) は，つぎのようになっている (添字は位置 i を表わす)：

> Sarah had managed to$_{160}$ open the world a little with her typewriter. That was$_{170}$ her work — typing. She did not type very quickly, and$_{180}$ so she had to work alone, and not in a$_{190}$ great office.
>
> The most successful of Sarah's battles with the$_{200}$ world was the arrangement that she made with Schulenberg's Home$_{210}$ Restaurant. The restaurant was next door to the old red-brick$_{220}$ building in which she had a room. ···

前半の場面はこの物語の主人公 Sarah の仕事に焦点があてられ，後半の場面は Schulenberg のレストランに焦点があてられているのがわかる．

[7] 図 6 の LCP は，方形窓を使用したものと考えてよい．

[8] 制限語彙を用いない原版は，2166 語 (786 語彙) である．

ここで注意すべきことは，テキストがどのように段落分けされているかに関係なく，LCP は場面境界をとらえることができることである．たとえば，$i=156$ の段落境界は，人間の直感による場面境界でも，LCP の極小点でもない．また $i=236$ は，LCP の共通場面境界かつ極小点の近くであるが，段落境界ではない．

しかしながら，LCP の極小点のいくつかは，共通場面境界と正確には一致していない．たとえば，$i=450$ の極小点は，$i=465$ の共通場面境界（かつ段落境界）とずれている．この付近の部分テキスト（$i=422\sim490$）は，つぎのようになっている:

> Both were satisfied with the agreement. Those who ate_{430} at Schulenberg's now knew what the food they were $eating_{440}$ was called, even if its nature sometimes puzzled them. And_{450} Sarah had food during a cold dull winter, which was_{460} the main thing with her.
>
> When the spring months $arrived_{470}$, it was not spring. Spring comes when it comes. The_{480} frozen snows of January still lay hard in the $streets_{490}$. ···

前半の場面は Sarah と Schulenberg との契約の結果に焦点があてられ，後半の場面は早春の寒さに焦点があてられている．問題となっているずれは，場面境界（つまり段落境界）前後の十数語が非常に結束している — ともに「寒さ」に関係している — ためであろう．

6 おわりに

本論文では，テキスト区画（とくに物語における場面）の境界をとらえるための新しい統計的な指標 LCP を提案した．LCP は，テキストの各位置について，その近傍の単語列の結束度を記録したものである．単語列の結束度は，それを構成する各単語相互の意味関係のつよさと定義され，英語辞書から規則的に構成された意味ネットワーク上の活性伝播によって計算される．

テキスト区画は意味的な一貫性をもち，そこに現われる単語はたがいに結束性によって結ばれる傾向がある．LCP は，このようなテキスト区画の性質を直接とらえる指標であり，人間の直観によるテキスト分割と比較した結果，LCP がテキスト区画の境界を推定するためのよい指標となることが確かめられた．

ここで提案したテキスト分割は，テキストの深い理解のためのボトムアップ的な情報を提供する．たとえば，つぎのような応用が考えられる:

- 照応や省略の解決:
  テキスト区画は，代名詞・指示詞などの参照対象を同定したり，省略された表現を復元するための，よい手がかり（制約・選好など）をもたらす．
- 要約の生成:
  テキスト区画から生成した活性パターンは，その区画の全体的・平均的な意味（**4.1** 参照）を表わし，要約をつくりだすための手がかりとなる．

今後は，ここで紹介した物語テキストだけでなく，ほかのジャンルのテキストについても，LCP の妥当性を検証する必要があると考えている．また，従来から提案されている手がかり語句などを併用し，より正確にテキスト区画の境界を推定できるように LCP を改良してゆくつもりである．

参考文献

[Grosz and Sidner, 1986] B. J. Grosz and C. L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12:175–204, 1986.

[Halliday and Hasan, 1976] M. A. K. Halliday and R. Hasan. *Cohesion in English*. Longman, Harlow, Essex, 1976.

[小嶋・古郡, 1993] 小嶋秀樹, 古郡廷治. 単語の意味的な類似度の計算. 電子情報通信学会技術研究報告, AI92-100:81–88, 1993.

[LDOCE, 1987] *Longman Dictionary of Contemporary English*. Longman, Harlow, Essex, new edition, 1987.

[Morris and Hirst, 1991] J. Morris and G. Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17:21–48, 1991.

[Schank, 1990] R. C. Schank. *Tell Me a Story: A New Look at Real and Artificial Memory*. Scribner, New York, 1990.

[Thornley, 1960] G. C. Thornley, editor. *British and American Short Stories*. Longman, Harlow, Essex, 1960. Longman Simplified English Series.

[Waltz and Pollack, 1985] D. L. Waltz and J. B. Pollack. Massively parallel parsing: A strongly interactive model of natural language interpretation. *Cognitive Science*, 9:51–74, 1985.

[West, 1953] M. West. *A General Service List of English Words*. Longman, Harlow, Essex, 1953.

[Youmans, 1991] G. Youmans. A new tool for discourse analysis: The vocabulary-management profile. *Language*, 67:763–789, 1991.